CLAIMS

What is claimed is:

1. A method for determining similarity between a first event set, the first event set comprising a first plurality of event types, and a second event set, the second event set comprising a second plurality of event types, the method comprising the steps of:

randomly mapping the first event set to a multidimensional vector-Q;

randomly mapping the second event set to a multidimensional query vector-q; and

determining similarity of the multidimensional vector-Q with the multidimensional query vector-q according to:

$||Q\text{-}q|| \leq SV$, where SV = a predetermined similarity value.

2. A method as in claim 1 wherein the step of randomly mapping the first event set to a multidimensional vector-Q further comprises the steps of:

determining a first weighting function; and

for each dimension j=1...k, where k is predetermined:

determining a first set of dimensional random variables for each event type within the first plurality of event types;

weighting each of the first set of dimensional random variables corresponding to each event type occurring within the first event set with the first weighting function; and

summing each of the weighted random variables.

3. A method as in claim 2 wherein the step of determining the first weighting function further comprises the step of determining a first linear weighting function.

4. A method as in claim 2 wherein the step of determining the first weighting function further comprises the step of determining a first exponential weighting function.

5. A method as in claim 2 wherein the step of determining the first set of dimensional random variables further comprises the step of determining a first set of normally distributed random variables.

6. A method as in claim 1 wherein the step of randomly mapping the first event set to a multidimensional vector-q further comprises the steps of:

determining a second weighting function; and

for each dimension $j_2=1 \ldots k_2$, where k_2 is predetermined:

determining a second set of dimensional random variables for each event type within the second plurality of event types;

weighting each of the second set of dimensional random variables corresponding to each event type occurring within the second event set with the second weighting function; and

summing each of the weighted random variables.

7. A method as in claim 6 wherein the step of determining the second weighting function further comprises the step of determining a second linear weighting function.

8. A method as in claim 6 wherein the step of determining the second weighting function further comprises the step of determining a second exponential weighting function.

9. A method as in claim 6 wherein the step of determining the second set of dimensional random variables further comprises the step of determining a second set of normally distributed random variables.

10. A method as in claim 1 wherein the first plurality of event types comprises:

at least one first dependent event type; and

at least one first independent event type.

11. A method as in claim 1 wherein the second plurality of event types comprises:

at least one second dependent event type; and

at least one second independent event type.

12. A method of finding a query-subset of events within an event set, the event set comprising a stream of ordered events, each ordered event corresponding to an event type e from a set of event types E, the method comprising the steps of:

dividing the stream of ordered events into a plurality of segments;

mapping each of the plurality of segments to a corresponding dimensional segment-vector;

mapping the query-subset of events to a dimensional query-vector; and

comparing the dimensional query-vector with at least one of the dimensional segment-vectors, and as a result of the comparison, making a determination whether the query-vector is similar to the compared segment-vector.

13. A method as in claim 12 wherein the step of mapping each of the plurality of segments to a corresponding dimensional segment-vector further comprises the steps of:

determining a first weighting function;

determining a random variable r(e,j) for each event type e for each dimension j=1...,k, where k is predetermined;

for each dimension j:

weighting the random variable r(e,j) for the event type corresponding to each event within the segment with the first weighting function to form a first group of dimension dependent weighted events; and

summing each of the dimension dependent weighted events within the first group.

14. A method as in claim 13 wherein the step of determining the first weighting function further comprises the step of determining a first linear weighting function.

15. A method as in claim 13 wherein the step of determining the first weighting function further comprises the step of determining a first exponential weighting function.

16. A method as in claim 13 wherein the step of determining the random variable r(e,j) further comprises the step of determining a normally distributed random variable.

17. A method as in claim 13 wherein the step of mapping the query-subset of events to a dimensional query-vector further comprises the steps of:

determining a second weighting function;

for each dimension j:

weighting the random variable r(e,j) for the event type corresponding to each event within the query-subset of events

with the second weighting function to form a second group of dimension dependent weighted events; and

summing each of the dimension dependent weighted events within the second group.

18. A method as in claim 17 wherein the step of determining the second weighting function further comprises the step of determining a second linear weighting function.

19. A method as in claim 17 wherein the step of determining the second weighting function further comprises the step of determining a second exponential weighting function.

20. A method for finding similar situations in sequences of events in a data flow, the method comprising the steps of:

mapping (4), using a random projection, slices (2) of a sequence (1) of events to multidimensional (k) vectors (5) and mapping a query sequence (3) of events to a multidimensional (k) query vector (7);

searching (6) among the vectors (5) for such multidimensional vectors (8) for which the distance from the query vector (7) is small;

for each slice corresponding to a found vector (8) verifying (9) its similarity to the query sequence (3); and

producing as a result a set of slices (10) that are similar to the query sequence (3).

21. A method according to claim 20, further comprising the steps of:

logging the sequences of events in the data flow over a period of time;

comparing the logged sequences of events with current data in the data flow; and

using the comparison results in predicting future events in the data flow.

22. A method according to claim 20 further comprising the steps of:

logging the sequences of events in the data flow over a period of time;

detecting regularities in the sequences of the data flow; and

using the detected regularities in producing individual services or interfaces.

23. A method according to claim 20, further comprising the steps of:

logging the sequences of events in the data flow over a period of time;

statistically analysing the events in the sequences of the data flow; and

using the statistical information in producing added value to applications and services.

24. A method according claim 20 wherein the sequences of events of the data flow comprise time series data, including sequences of events with their corresponding occurrence times in a telecommunications network.

25. A method according claim 20 wherein the sequences of events of the data flow comprise data arising from mobile services with request being made by mobile subscribers or from user interface studies where users manipulate a device through an interface or data arising from financial or scientific applications, such as stock price indices, the volume of product sales, telecommunication data, medical signals, audio data or environmental measurement sequences.

26. A method according to claim 20 wherein the sequences of events of the data flow comprise ordered data objects or events with a defined relative order, the relative order of the data objects or events defined as their occurrence time.

27. A method according claim 20, wherein the slices comprise subsets of the sequence (1), defined as the subsequence of events of (1) occurring in a defined time interval.

28. A method according claim 20 wherein the query sequence (3) comprises a sequence of events, defined as a sequences of events for which the occurrence times are within a defined time interval.

29. A method according to claim 20 wherein the mapping using random projections comprises a mathematical random projection from sequences of events to a multidimensional (k) space, the random projection having the property of small expected distance between vectors mapped from similar sequences and a larger expected distance between vectors mapped from dissimilar sequences.

30. A method according to claim 20 wherein the searching (6) further comprises the step of linear searching.

31. A method according to claim 20 wherein the searching (6) further comprises the step of advanced data structure searching.

32. A method according to claim 20 wherein the distance between vectors comprises the norm of the difference between the vectors.

33. A method according to claim 20, wherein the step of the verification (9) comprises the step of applying edit distance computation for establishing similarity between slices and the query sequence (3).

34. A system for finding a query-set of events within a master-set of observed events, wherein the events belong to an event set E, the system comprising:

a k-dimension random variable generator for generating random variables for each event within the event set E for each j=1...k dimensions, where k is predetermined;

an observed event segmenter for segmenting the master-set of observed events to produce a plurality of observed event segments, $(d_{11},s_{11}\ldots d_{1m},s_{1m})\ldots$ $(d_{h1},s_{h1}\ldots d_{hm},s_{hm})$, where d=segmented observed event, and s= a parameter associated with the observed event;

a weighting function generator for generating a weighting function;

an observed event vectorizer for vectorizing each observed event segment $(d_{11},s_{11}\ldots d_{1m},s_{1m})\ldots$ $(d_{h1},s_{h1}\ldots d_{hm},s_{hm})$ according to the weighting function and the dimensional random variable corresponding to the segmented observed event d and dimension k;

a query event vectorizer for vectorizing the query set of events ($Qe_{1,}Qs_1 \ldots Qe_wQs_{w,}$) according to the weighting function, the dimensional random variable corresponding to the query event Qe, and dimension k, wherein Qe = a query event, Qs = a parameter associated with the query event Qe; and

a comparator for comparing the vectorized query-set of events with each vectorized observed event segment and generating the observed event segment in accordance with the comparison results and predetermined similarity factors.

35. A system as in claim 34 wherein the k-dimension random variable generator further comprises a k-dimension normally distributed random variable.

36. A system as in claim 34 wherein the master-set of observed events comprises a master-set of time ordered events.

37. A system as in claim 34 wherein the s parameter comprises a time parameter associated with the observed event d.

38. A system as in claim 34 wherein the weighting function generator comprises a linear weighting function generator.

39. A system as in claim 34 wherein the weighting function generator comprises an exponential weighting function generator.

40. A system as in claim 34 wherein the weighting function generator comprises a look-up-table.

41. A system as in claim 34 wherein the system further comprises an observed event data preconditioner for segmenting observed event data into

at least one dependent observed event data segment or at least one independent observed event data segment.

42. A system as in claim 34 wherein the system further comprises a query event data preconditioner for segmenting query event data into at least one dependent query event data or at least one independent query event data.

43. A system as in claim 34 wherein the system further comprises a query event data preconditioner for statistically analyzing the query event data and segmenting the query event data according to the results of the statistically analyzed query event data..

44. A system as in claim 43 wherein the system further comprises a observed event data preconditioner for preconditioning the observed event data according to the results of the statistically analyzed query event data.

45. A system as in claim 34 wherein the system further comprises an observed event data preconditioner for statistically analyzing the observed event data and segmenting the observed event data according to the results of the statistically analyzed observed event data.

46. A system as in claim 34 wherein the parameter Qs associated with the query event Qe further comprises a time parameter.

47. A system as in claim 46 wherein the query set of events further comprises a time ordered set of events according to the associated parameter Qs.

48. A program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform method steps determining similarity between a first event set, the first event set comprising a first plurality of event types, and a second event set, the second event set

comprising a second plurality of event types, the method comprising the steps of:

mapping the first event set to a multidimensional vector-Q;

mapping the second event set to a multidimensional query vector-q; and

determining similarity of the multidimensional vector-Q with the multidimensional query vector-q according to:

$||Q-q|| \leq SV$, where SV = a predetermined similarity value.